University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

01 University of Plymouth Research Outputs University of Plymouth Research Outputs

2015-06

How clumpy is my image?

Hutt, H

http://hdl.handle.net/10026.1/9340

10.1007/s00500-014-1303-z
Soft Computing
Springer Science and Business Media LLC

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Soft Computing manuscript No.
(will be inserted by the editor)

How Clumpy is my Image?

Scoring in Crowdsourced Annotation Tasks

Hugo Hutt · Richard Everson · Murray Grant · John Love · George Littlejohn

Received: date / Accepted: date

Abstract The use of citizen science to obtain annotations from multiple annotators has been shown to be an effective method for annotating datasets in which computational methods alone are not feasible. The way in which the annotations are obtained is an important consideration which affects the quality of the resulting consensus annotation. In this paper, we examine three separate approaches to obtaining consensus scores for instances rather than merely binary classifications. To obtain a consensus score annotators were asked to make annotations in one of three paradigms: classification, scoring and ranking. A web-based citizen science experiment is described which implements the three approaches as crowdsourced annotation tasks. The tasks are evaluated in relation to the accuracy and agreement among the participants using both simulated and real-world data from the experiment. The results show a clear difference in performance between the three tasks, with the ranking task obtaining the highest accuracy and agreement among the participants. We show how a simple evolutionary optimiser may be used to improve the performance by reweighting the importance of annotators.

Hugo Hutt, Richard Everson
Computer Science, The University of Exeter, Exeter, UK
E-mail: {hwh202, R.M.Everson}@exeter.ac.uk

Murray Grant, John Love, George Littlejohn
Biosciences, The University of Exeter, Exeter, UK
E-mail: {M.R.Grant, J.Love, G.R.Littlejohn}@exeter.ac.uk

1 Introduction

The increasing digitisation of science has dramatically reduced the cost of generating large amounts of data in a relatively short amount of time. However, the analysis and labelling of this data often requires human participation. In fact, annotating large datasets to provide ground-truth instances has become one of the major bottlenecks for developing effective supervised machine learning models which can generate new predictions [Raykar and Yu, 2012]. Alternatives to purely computational approaches are therefore required in order to obtain the annotations.

Citizen science seeks to elicit the help of non-experts to address scientific problems by using *crowdsourcing* [Doan et al., 2011]. Often this takes the form of an online annotation task in which the collective efforts of many individual participants are used to arrive at estimates of the consensus annotations. Recently, a number of citizen science projects have shown effectiveness in using crowdsourcing approaches to acquire annotated datasets which can then be used to guide computational approaches [Whitehill et al., 2009, Fortson et al., 2012, Parent and Eskenazi, 2010]. Annotations gathered from citizen science experiments can result in valuable training data for machine learning models, while also providing insights into the behaviour of the participants. In addition, there are a number of interesting theoretical problems surrounding citizen science as a result of the different degrees of accuracy associated with the partic-

Author contributions: HH wrote the software, recruited participants, ran the experiment and analysed the data; RE supervised the computer science aspects; GL conducted the imaging experiments and provided biological expertise; MG provided biological materials and expertise in infection biology; JL provided biological expertise. HH, RE and GL wrote the paper.

ipants and the uncertainty inherent in the data [Raykar et al., 2010].

With the increasing number of online projects there is a corresponding need to investigate how crowdsourcing tasks should be presented to the participants [Heer and Bostock, 2010, Parent and Eskenazi, 2010, Snow et al., 2008]. The effect that different types of annotation tasks have on the performance and consensus of the participants is an important, but largely unexplored topic. The choice of task is an essential consideration when using crowdsourcing to gather annotations, as it determines to a significant extent the quality of the resulting data. A common and relatively well-understood task is classification in which annotators are asked to assign instances to one of a number of discrete classes. Since the classes are predefined through criteria determining membership of each class, the annotator's task is conceptually straightforward, even if determining to which class an instance belongs is difficult.

In contrast, assigning a *score* to an instance is more difficult because individual annotators may use the range of scores differently and may judge the linearity of the scale differently. The goal of this paper is to investigate a number of separate approaches to obtaining score annotations from experimental participants and to examine their effectiveness. We describe a web-based citizen science experiment involving the annotation of microscopy images of plant cells during bacterial infection. Briefly, the goal is to assess the degree of "clumpiness" of each image. This task therefore differs from the more common classification task, in which the annotator is asked to assign an object to discrete categories, because "clumpiness" is a continuous quantity. Three separate paradigms are used to obtain the image annotations and in this paper we assess their efficacy and means of deriving a consensus score from the annotations. The approaches are evaluated on both simulated and real-world data from the experiment and a comparison is made between the different tasks. In particular, the influence of the task type on the overall performance and consensus of the annotators is examined. The annotation of the microscopy images is a very challenging problem for current image processing techniques, which makes it a good candidate for a citizen science project.

The rest of the paper is organised as follows. Section 2 is a description of the problem. Section 3 describes the citizen science experiment, including the user statistics for each of the tasks. Section 4 outlines methods for evaluating different annotation tasks. Section 5 describes the simulation setup used to model annotators under the different tasks. Section 6 presents the empirical results from the simulated and experimental data. Section 7 describes how the estimates of individual annotators can be reweighted using an evolutionary optimiser to obtain more accurate results. Section 8 concludes the paper.

2 Description of Problem

2.1 Learning from Multiple Annotators

In a typical annotation task, there is a set of N instances $\mathbf{x} = \{x_1, \ldots, x_N\}$ whose true annotations are unknown. Each instance $x_i \in \mathbf{x}$ is then assigned an annotation by R annotators, resulting in a set of estimates $\{y_i^1, \ldots, y_i^R\}$ of the true annotation. Given these multiple annotations, the goal is to arrive at accurate consensus estimates $\mathbf{y} = \{y_1, \ldots, y_N\}$ for each of the N instances.

One simple and often used technique for obtaining consensus estimates from multiple annotators is majority voting [Raykar et al., 2010]. For binary classification, the majority vote estimate of an instance x_i is defined as

$$y_i = \begin{cases} 1 & \text{if} \quad \frac{1}{R}\sum_{j=1}^{R} y_i^j \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $y_i^j \in \{0, 1\}$ is the annotation assigned to instance x_i by annotator j. For simplicity of notation we assume that each instance is annotated by the same number, R, of annotators, although in practice R is often different for each instance. Majority voting can be extended to scores, where each instance is assigned the mean of the annotators' scores:

$$y_i = \frac{1}{R}\sum_{j=1}^{R} y_i^j. \tag{2}$$

If the scores are made on an integer scale (*e.g.,* a five-point scale: $y_i^j \in \{1, 2, 3, 4, 5\}$), the estimate y_i can then be rounded to the nearest score on the scale.

Ideally, annotators with higher accuracy should be given more weight when estimating the consensus, while the influence of poor quality annotators should be decreased or removed entirely. A major limitation of standard majority voting is that it assumes all annotators are equally reliable, meaning that its effectiveness is dependent on the overall quality of the annotators [Raykar et al., 2010]. Given an estimate of an annotator's performance, we can introduce an additional weighting term to the vote to account for the variation in quality among the annotators. Let ϵ_j be the error rate of annotator j on some subset of the instances for which the true annotations are known. The standard majority vote for

classification can be replaced by

$$y_i = \begin{cases} 1 & \text{if} \quad \frac{1}{R}\sum_{j=1}^{R} |y_i^j - \epsilon_j| \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

which penalises annotators with high error rates and assigns greater weight to the estimates of accurate annotators. If there is no known standard on which to evaluate the annotators, more sophisticated techniques are required in order to account for the differences in annotator quality, such as those proposed in [Whitehill et al., 2009, Raykar et al., 2010, Raykar and Yu, 2012].

In Section 7 we describe how annotator weighting parameters can be optimised to improve the accuracy of the estimates.

2.2 Obtaining Scores from Annotators

The annotations collected from the experiment indicate the degree of "clumpiness" present in the microscopy images. This notion of clumpiness is continuous in nature, with scores potentially falling within an indefinite range. Unlike classification tasks, which involve assignment to predefined categories, ways of assessing a score are less well explored. We therefore asked annotators to perform three different kinds of task in order to elicit a consensus score. Each individual annotator was randomly assigned to one of these tasks and did not annotate images using the other two. The following is a description of the three kinds of annotation investigated in this paper: classification, scoring and ranking.

The classification task divides the range of scores into two (*not clumpy* and *clumpy*) and requires the annotators to assign binary scores $\{0, 1\}$ to the instances:

$$y_i^j \in \{0, 1\} \quad \forall x_i \in \mathbf{x}. \tag{4}$$

A consensus classification is then obtained by majority voting (1). In addition, the proportion of 1 annotation is assigned (2). Clearly, this score can be interpreted as the probability that the instance belongs to either class. To obtain the maximum amount of information the class boundary should be placed so that approximately half of the instances fall in either class. However, while this task is conceptually straightforward for annotator, they may find it difficult to assign instances close to the artificially-imposed division between the classes. Furthermore, the extreme "quantisation" of the continuous scale into just two categories inevitably discards information about degree which is only recovered after many annotations have been made.

For the scoring task, the annotators directly assign scores in a pre-determined range. Although in principle an indefinitely fine scale could be employed, in our experiments a seven-point integer scale was used:

$$y_i^j \in \{1, \ldots, 7\} \quad \forall x_i \in \mathbf{x} \tag{5}$$

A fairly coarse integer scale, like this, relieves annotators of feeling that they have to make very fine distinctions, while allowing them to distinguish between *very clumpy* and *quite clumpy*, etc. Nonetheless, even when furnished with examples, annotators may not use the full range of the scale and, of course, may assign different scores based on their prejudices and the particular instances that they have seen previously. Clearly, a consensus score is easily given by the mean of the annotators' scores (2).

For the ranking task, annotators are required to rank-order subsets of the instances according to whatever quantity is being assessed. This results in a set of ordered relations and we write $(x_i \prec_j x_k)$ to indicate that x_i has been assessed to have a lower score than x_k by annotator j. Although probabilistic models for inferring ranks for partial information can be constructed [e.g. Lebanon and Lafferty, 2002], we use a straightforward method for determining a consensus score as follows. Consider the specific case in which each instance is ranked either higher or lower than one other instance. From these binary rankings, a score is derived for each of the instances. Let

$$\mathcal{R}_{x_i} = \{x_k \in \mathbf{x} \mid (x_k \prec_j x_i) \forall j\} \tag{6}$$

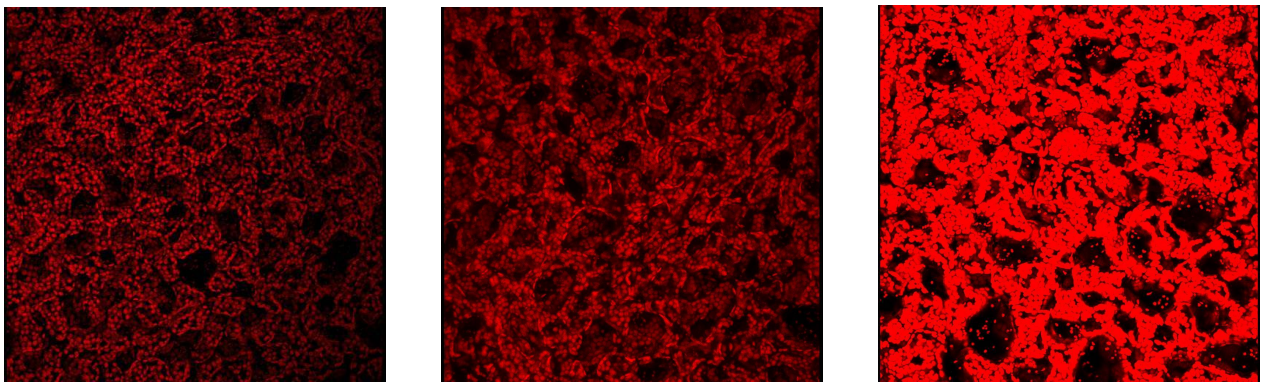
be the set of instances ranked lower than x_i by any annotator. Also let

$$\mathcal{T}_{x_i} = \{x_k \in \mathbf{x} \mid (x_k \prec_j x_i) \vee (x_i \prec_j x_k) \forall j\} \tag{7}$$

be the instances ranked either lower or higher than x_i. The consensus score for x_i is then

$$y_i = \frac{|\mathcal{R}_{x_i}|}{|\mathcal{T}_{x_i}|} \tag{8}$$

so that $0 < y_i \leq 1$. Clearly, instances that are consistently ranked above other instances will obtain high scores and *vice versa*. In the experiment we describe below, annotators were asked to rank order groups of three images from least clumpy to most clumpy. This ranking was decomposed into the three implied binary relations and used as described above. The advantage of the ranking task is that annotators find it relatively easy to compare instances and agree on an ordering even if they disagree on a precise score or even to which pair of classes an instance belongs. Unlike the classification and scoring tasks there is no need for the annotator to refer back to a set of fiducial instances for calibration.



(a) Consensus Score: 0.14 (b) Consensus Score: 0.39 (c) Consensus Score: 0.85

Fig. 1 Examples of the chloroplast images used for the experiment. Also shown are the consensus scores from the ranking task, which range from 0 to 1.

3 Description of Experiment

In order to assess the different approaches to obtaining the instance annotations outlined above, we describe here a web-based citizen science experiment involving the annotation of plant cell images according to their "clumpiness". [1]

We first describe how the image dataset used for the experiment was acquired. The microscopy images obtained for the experiment show perfluorocarbon-mounted [Littlejohn et al., 2010] leaves of the model plant *Arabidopsis thaliana* (Col-0 ecotype) obtained using a Zeiss 510Meta Laser Scanning Confocal Microscope equipped with a 40x oil immersion lens. Chlorophyll was imaged by Excitation at 488 nm and Emitted Fluorescence was collected with a LP615 nm filter. So-called "Z-stacks", consisting of 75 parallel planar images with an interplanar separation of 1 μm were collected during a timecourse comparing infection with the phytopathogenic bacterium *Pseudomonas syringae* pv. tomato strain DC-3000 to a mock inoculation using infection conditions previously described [Truman et al., 2006].

When leaves were infiltrated with the virulent bacteria, it was noticed that during the timecourse, chloroplasts tended to clump together within the cell. The goal of the citizen science experiment was to determine the clumpiness of each image by deriving a consensus score from the individual annotations provided by multiple annotators. In addition, the resulting labelled dataset can potentially be used as training data for supervised learning algorithms.

The participants were shown chloroplast-only 3D maximum projections of confocal z-stacks, comprising 19 projections turning round the z-axis with a first angle of 45° and a difference angle of −5°. Participants were free to rotate these projections, static examples of which are shown in Figure 1. When first registering on the site, a tutorial page was displayed to the participants which included some example images with known scores. Note that this was the only training provided to participants and no feedback was given during the experiment.

There were three tasks associated with the experiment, which can be viewed as implementations of the classification, scoring and ranking approaches outlined in Section 2. The classification task involved classifying the images as either "clumpy" or "not clumpy" by selecting the appropriate button. For the scoring task, the participants used a slider bar to specify a "clumpiness" score from 1 to 7 for the images. Finally, the ranking task required the annotators to order groups of three images left to right, from least clumpy to most clumpy. This was achieved by dragging the images into the desired position. The participants were assigned one of the three tasks randomly on registering with the website and only annotated images in a single paradigm. There was no limit to the number of annotations, with each participant free to annotate as many of the images as they chose.

Table 1 summarises the annotation statistics for each of the tasks. Note that each individual ranking provides information about three images, whereas information on only a single image is obtained from each scoring or classification.

In addition to the initial dataset of 64 images, we also created a second set by rotating the original images by 180 degrees. This was carried out to provide a means of evaluating the reliability of annotators, as the degree of clumpiness for the original and rotated images will be identical.

[1] The URL of the site is `http://www.clumpy.ex.ac.uk` which remains active at the time of writing.

Table 1 Summary of annotation statistics from the experiment. The table shows for each task the number of annotators, the number of annotations performed and the average number of annotations per annotator.

Task	Annotators	Annotations	Avg PA
Classification	76	3410	45
Scoring	74	3709	50
Ranking	77	1605	21
Total	**227**	**8724**	**39**

In order to evaluate the accuracy of annotators from the three tasks, a *gold standard* was selected consisting of seven images annotated by an expert. These images were chosen so as to include the full range of possible scores (i.e. from 1 to 7). A randomly selected image (or group) from the gold standard was displayed to the participants at regular intervals, enabling an "Expert vs. Participant" measure of accuracy on the images.

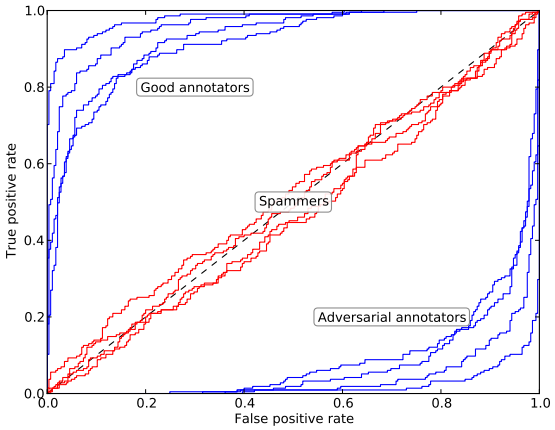


Fig. 2 The good and adversarial annotators are above and below the diagonal of the ROC, respectively. Spammers are close to the diagonal of the ROC, assigning the scores at random.

4 Evaluating Annotation Tasks

In order to evaluate the accuracy of both individual annotators and the consensus scores on a common footing we view all three tasks in a classification framework, which allows evaluation of performance using the commonly used Receiver Operating Characteristic (ROC) curves. Given some instances (the gold standard images) whose true classifications are known, the ROC curve displays the true positive rate versus the false positive rate as the decision threshold is varied and the area under the ROC curve (AUC) measures the overall ability of a classifier to separate two classes [Fawcett, 2006].

Clearly application of the ROC methodology to the classification task is straightforward: the performance of a consensus classifier can be evaluated on the gold standard images and the performance of individual annotators can be evaluated against either the gold standard or the consensus classification. Annotators can be viewed as belonging to one of three general quality classes, depending on their classification performance [Raykar and Yu, 2012]. Example ROC curves for annotators of different quality are shown in Figure 2 using simulated data.

A *good* annotator's ROC curve lies above the diagonal of the plot, indicating that they consistently make correct annotations. The proportion of good annotators in the population and their overall level of performance depends on a number of factors, such as the individual difficulty of the instances, the duration of the task, as well as the accuracy and reliability of the annotators.

An *adversarial* annotator's ROC curve lies below the diagonal of the ROC plot. These annotators are the mirror image of the good annotators on the ROC plot, assigning incorrect annotations to the instances. An important point to note is that although adversarial annotators are inaccurate and assign incorrect annotations, they do so consistently. This means that if they can be detected in the population and have their annotations "flipped", they still have discriminatory power [Raykar and Yu, 2012].

Finally, a *spammer* is an annotator who assigns annotations at random [Raykar and Yu, 2012]. For binary classification, this corresponds to the situation in which the annotator is close to the diagonal of the ROC plot, as shown in Figure 2. Annotators close to the diagonal of the ROC provide no useful discriminatory power and their annotations should be ignored or removed if they are detected in the population.

Although annotators tend to fall into one of the three classes, the distinction is not always easy to make. For example, an annotator may start off as random or adversarial, but improve their accuracy as they are exposed to more instances. Conversely, an annotator's accuracy can also decrease over time.

In addition to evaluating the accuracy, a number of other properties of the annotators were assessed when comparing the three tasks. We measured how strongly the annotators were correlated with each other and how reliable they were in maintaining their accuracy for the duration of the task. The results are presented in Section 6.

Annotations in the ranking class are easily cast in a classification framework by considering the binary relations that result from ordering the instances. We denote a ranking as correct if the two instances involved are placed in their true order (obtained from the gold standard or consensus) and incorrect if not.

Finally, the scoring task is cast in a classification framework denoting a score as correct if it and the true score are both greater than or equal to 4 (the middle of the available range) or if both are less than or equal to 4; otherwise the score is deemed incorrect. Although other more sensitive loss functions might be used in this context, this provides a common framework for evaluating performance in all three tasks.

5 Simulation

In order to investigate the annotation tasks under various conditions, simulated data was generated to model annotators with different degrees of accuracy and performance. In a simple model, whether each annotator correctly annotates an instance could be modelled by a draw from a Bernoulli distribution $\text{Be}(\pi)$ with an annotator-specific probability π. To provide a richer model, accounting for the variability in each annotator's performance, the probability π was modelled by a beta distribution [Gelman et al., 2013]. The beta distribution is defined by

$$p(\pi;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{9}$$

where the two parameters $\alpha > 0$ and $\beta > 0$ control the mean and spread of the distribution. The values of α and β were assumed to be annotator specific.

6 Empirical Results

6.1 Accuracy of Annotators

Estimates of the annotator parameters for the beta distribution were obtained using maximum likelihood and the observed accuracy on the gold standard. The beta distributions using these parameter estimates are shown for each of the annotators using each of the different annotation paradigms in Figure 3. The ranking task distributions tend to be more sharply peaked compared to the other two tasks, with mean $\mu = 0.75$ and standard deviation $\sigma = 0.09$ for the mean distribution. This indicates that there was less variation in accuracy among the annotators. The participants from the classification ($\mu = 0.75$, $\sigma = 0.15$) and scoring ($\mu = 0.68$, $\sigma = 0.17$) tasks tended to be less reliable in their estimates, obtaining a wider range of accuracies.

Using these parameter estimates, ROC curves of simulated annotators corresponding to each of the actual annotators were obtained for the different tasks. Figure 4 shows the results; note that a small jitter was added to the curves to aid visualisation, as annotators with very high accuracy tend to be concentrated around the top-left corner of the ROC plots. The ROC curves derived from the mean parameter estimates are shown in bold. A number of annotators were close to the diagonal of the ROC, indicating the presence of spammers in the population. Adversarial annotators can also be clearly identified in each of the tasks, as shown by the curves lying below the diagonal of the ROC. The mean curves are seen to obtain good performance, with the ranking task in particular being near-optimal.

The accuracy of the annotators was also evaluated in relation to the consensus (majority vote) annotations for the images. Figure 5 shows the consensus accuracy for annotators versus the number of image annotations they made. As can be seen, the consensus accuracy of the annotators tended to remain stable, with no large increase or decrease in the accuracy as more annotations were made. Figure 5 also shows the number of annotators who made a given number of annotations in each of the three paradigms. Here it is clear that ranking annotators made relatively fewer annotations than scorers or classifiers, but note that, although each ranking task is more time-consuming, information on three images is obtained from each annotation.

In terms of the overall accuracy, the ranking task obtained the best performance. The greater proportion of annotators with high accuracy was reflected in the performance of the consensus estimates.

6.2 Inter-Annotator Agreement

The inter-annotator agreement provides a measure of the consensus among multiple annotators, which enables a comparison between different annotation tasks in terms of the agreement among the participants. We used the Spearman rank correlation [Lehmann, 2006] for the comparison, which is a non-parametric statistic measuring the strength of association between two sets of data. Let $\{y_i^j\}$ and $\{y_i^k\}$, $i = 1,\ldots,R$, be the sets of image annotations in common between annotators j and k. The Spearman correlation between the two annotators is then defined as

$$\rho_{jk} = 1 - \frac{6\sum_{i=1}^{R}(\sigma_i^j - \sigma_i^k)^2}{R(R^2-1)} \tag{10}$$

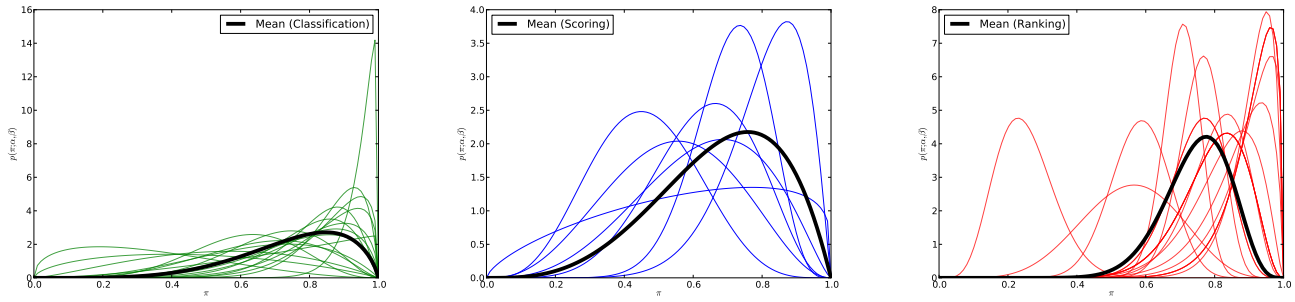


Fig. 3 Annotator beta distributions for the classification, scoring and ranking tasks, respectively. Shown in bold are the distributions for the mean parameter estimates.

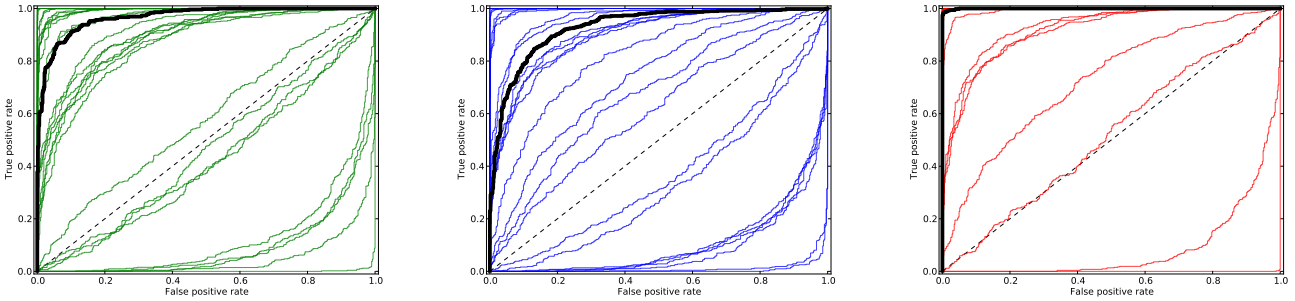


Fig. 4 ROC curves of simulated annotators using the parameter estimates from the experiment. From left to right, the plot shows the results from the classification, scoring and ranking tasks. The ROC curves for the mean parameter estimates are shown in bold. A small jitter has been added to these curves to separate them for visualisation.

where σ_i^j is the rank of y_i^j in the set of annotations $\{y_i^j\}$. By evaluating the correlation between each pair of annotators, we can compute the average agreement for individual annotators. The agreement can also be used to distinguish between the different types of annotators described in Section 4. Adversarial annotators will tend to have negative agreement with the good annotators, whereas spammers (random annotators) will tend to have an average agreement near to 0.

The mean inter-annotator agreement was obtained for each annotator by computing their average Spearman correlation with all other annotators assigned the same task. From Figure 6a it can be seen that there were also relatively few negatively correlated annotators, with the large majority obtaining positive average correlations. However, it can be seen that participants carrying out the classification task had significantly lower agreement than those from the ranking and scoring tasks. In addition there were more spammers (with correlation close to 0) for the classification task, which is evident from the individual ROC curves shown in Figure 4.

Figure 6b is the result of bootstrapping on the set of mean inter-annotator agreements from each task. It shows 1000 bootstrapped sample estimates of the mean and standard deviation for each task. The separation of the classification task from the other two is clear, with the annotators showing only small variations in agreement. The scoring and ranking tasks on the other hand show more variation in addition to a higher mean agreement.

As the results show, the ranking task obtained the highest level of agreement among the annotators. The participants from the classification task obtained significantly lower agreement. This is partly to be expected due to the nature of the task, as there is no notion of the degree to which classifiers agree on an instance, only whether they agree or disagree.

6.3 Reliability of Annotators

In order to test the reliability of the annotators, we calculated their accuracy in relation to the consensus scores on both the original and rotated images. Analysis of variance (ANOVA) [Bailey, 2008] was then used to compare the consensus accuracy on the original and rotated images. This gives an idea of how consistently

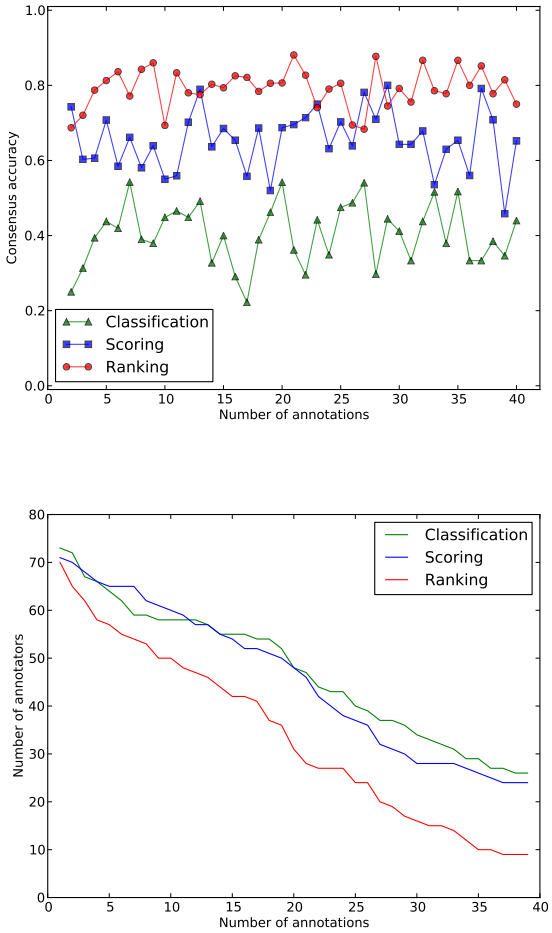


Fig. 5 The top figure shows the mean consensus accuracy of annotators plotted versus the number of annotations they made. The bottom figure shows how many annotators made a particular number of annotations.

Table 2 Results from ANOVA on the consensus accuracy for the original and rotated images. The table reports the F-ratios and p-values obtained for each task.

Task	$F(1, 62)$	p
Classification	0.554	0.458
Scoring	0.36	0.85
Ranking	0.003	0.953

the annotators maintained their accuracy throughout the duration of the tasks. The results are shown in Table 2.

None of the tasks showed a statistically significant difference between the accuracies on the original and rotated images. The ranking task showed a particularly strong similarity between the two sets of accuracies, indicating that the annotators were reliable in estimating the degree of clumpiness present in the images.

An indication of the correspondence between annotators from each of the tasks is seen in Figure 7. The plot shows the combined scores on both the original and rotated images, sorted in increasing order and translated to a common scale. Note that all consensus scores from all three tasks agree on the position in which images should be placed in order of clumpiness. The scores at the two endpoints of the plot show less of a divergence between the three tasks than those around the middle. This suggests that the annotators had more difficulty in determining the degree of clumpiness in the middle range, compared to the more obviously clumpy/not-clumpy images at the two ends of the scale.

6.4 Required Number of Annotators

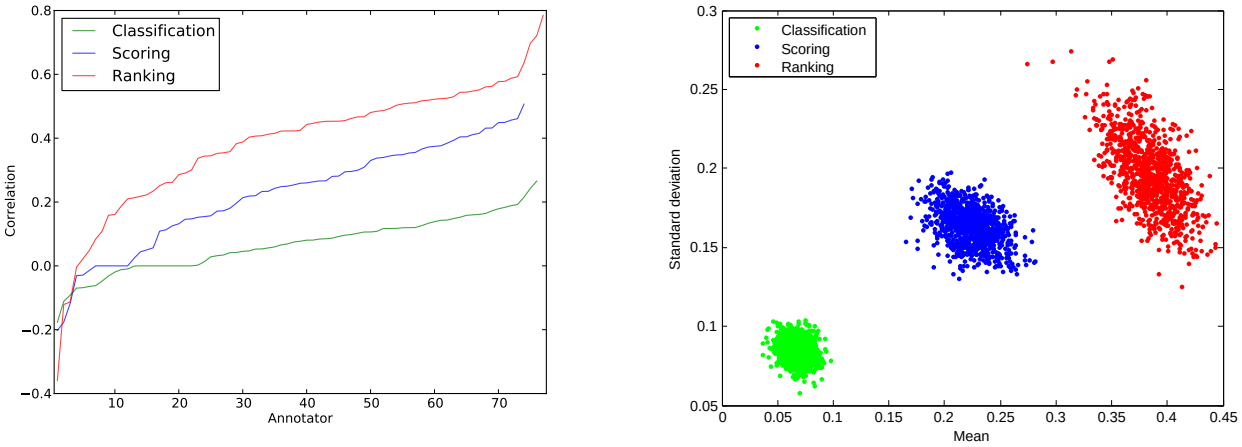
While the accuracy of the overall population of annotators has been considered, in this section we investigate the minimum number of annotators required in order to obtain accurate consensus estimates. By repeatedly sampling at random from the population, the accuracy was evaluated for different numbers of annotators by taking the consensus over those annotators and comparing with the gold standard; 10000 samples were used here. Figure 8 shows plots of the obtained accuracy for each of the tasks.

Comparing the results reveals that the increase in accuracy is more apparent during approximately the first 20 annotations, after which the accuracy begins to converge. The scoring task showed no improvement after the number of annotators increased beyond 20, whereas the classification and ranking tasks continued to show a gradual improvement as the sample size was increased past 50 annotators. The ranking task was ultimately able to reach a higher mean accuracy than the other two tasks.

7 Optimising Weighting Parameters for Consensus Annotations

7.1 Description of Model

An important consideration when estimating the consensus annotations is how to weight the annotators according to their quality. The results reported thus far have given equal weight to annotators, but it is clear from the ROC curves shown in Figure 4 that both spammers and adversarial annotators are present. Using the consensus accuracy on the gold standard as the objective, we optimised weighting parameters for the annotators using an evolutionary algorithm. The



(a) (b)

Fig. 6 (a) The mean inter-annotator agreement of annotators from each task, plotted as the mean correlation of each annotator with the others. For each task the annotators are ordered by correlation. (b) Bootstrapped sample estimates of the mean and standard deviation of the inter-annotator agreements. The plot shows 1000 bootstrapped sample estimates for each task.

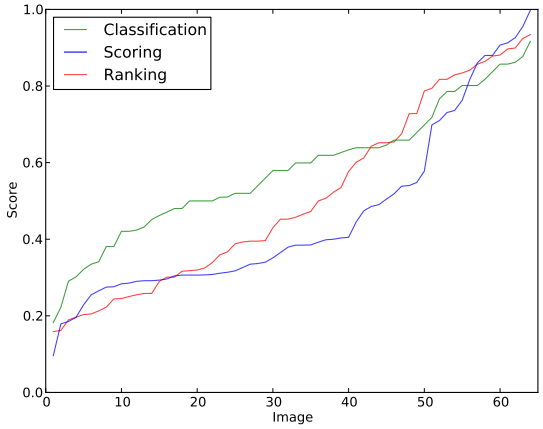


Fig. 7 Combined image scores (original and rotated) from each of the tasks, in increasing order and translated to a common scale. The classification scores are the fraction of positive classifications for an image. The values for the scoring task are the mean scores for the images. Finally, the ranking task values are the derived scores described in Section 2.2.

weighted consensus estimate of an instance was obtained using

$$y_i = \sum_{j=1}^{R} w_j y_i^j \tag{11}$$

where w_j is the weighting parameter for annotator j. The weights themselves constrained so that $w_j \in [-1, 1]$ and

$$\lambda \sum_{j=1}^{R} |w_j| = 1 \tag{12}$$

This normalises the consensus estimates and allows adversarial and spamming annotators to be assigned negative and zero weights respectively. In a similar manner to the LASSO method [Tibshirani, 1996] this l_1-norm penalty acts to promote sparseness. The constant λ controls the degree of sparsity, with larger values leading to more annotators being assigned weights close to 0.

Optimum weights were found by minimising the average error on the gold standard instances penalised by (12) using a straightforward elitist evolutionary algorithm similar to the well-known Pareto Archived Evolution Strategy (PAES) [Knowles and Corne, 1999], but without explicit diversity control. At each generation the algorithm randomly perturbed the annotator weights of selected members of a population of weights, retaining the best members of the population for the succeeding generation. Algorithm 1 shows pseudocode for the procedure. The value of λ was chosen experimentally for each task by evaluating the optimisation within a specified range and selecting the value which resulted

Algorithm 1 Weighted consensus optimisation

Require: Annotations for each $x_i \in \mathbf{x}$ from R annotators
1: Initialise archive E of weights
2: **repeat**
3: Select weights $\mathbf{w}$ from E
4: Perturb current weights $\mathbf{w} \rightarrow \mathbf{w}'$
5: Evaluate accuracy of (11) with $\mathbf{w}'$
6: **if** accuracy improves **then**
7: Replace $\mathbf{w}$ with $\mathbf{w}'$ in E
8: **end if**
9: **end**
Ensure: Final optimised weights in $\mathbf{w}$

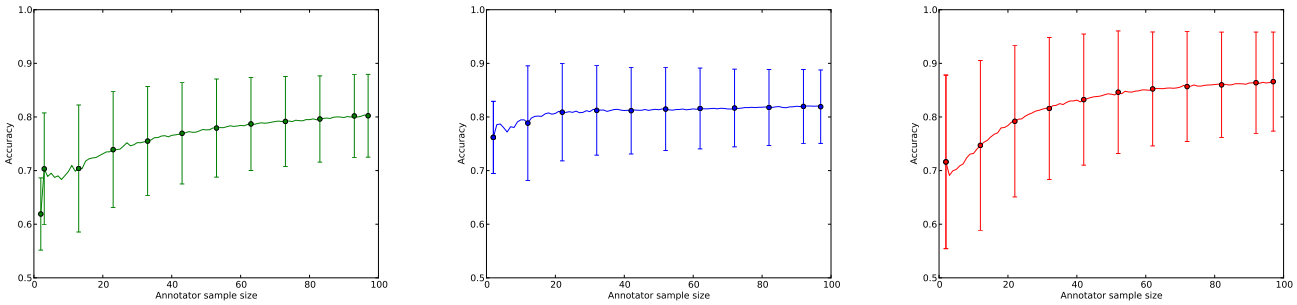


Fig. 8 Mean consensus accuracy of the classification, scoring and ranking tasks as a function of the number of annotators. For each sample size, the figures show the mean accuracy of 10000 random samples of that size from the population of annotators. One standard deviation error bars are also shown.

Table 3 Results from LOO cross-validation on the gold standard. The table shows the average accuracy obtained on both the training and test data.

Task	Training Acc	Test Acc
Classification	1.0	1.0
Scoring	1.0	0.86
Ranking	0.95	1.0

in the best overall performance on the gold standard instances. Typically, λ ranged from 0.01 to 0.075.

7.2 Model Validation

In order to validate the model, leave-one-out (LOO) testing was carried out to assess how well the model generalises to unseen data. On each training run, one of the gold standard instances was held out and the weights were optimised on the remaining instances; the accuracy was then evaluated on the held out instance using the optimised weights to obtain the test accuracy. This was repeated for all of the instances in the gold standard. Table 3 shows the training and test accuracy for each task averaged over the independent runs. The λ constant was set to 0.075, 0.025 and 0.015 for the classification, scoring and ranking tasks, respectively.

The consensus estimates from all three of the tasks were able to obtain high training and test accuracy on the gold standard using the optimised weighting parameters. For comparison, the standard (equally-weighted) consensus accuracy was 0.86, 0.57 and 0.86 for the classification, scoring and ranking tasks, respectively. Given that the annotators tended to be reliable in maintaining their accuracy, this means a significant improvement in the overall quality of the consensus estimates can be expected.

The top row of Figure 9 shows plots of the optimised annotator weightings for each of the tasks, averaged over all training runs of the cross-validation. It can be seen that annotators were assigned a range of weights, indicating differences in quality. The annotators assigned weights close to 0 can be assumed to include spammers, while those assigned negative weights are more adversarial. It is interesting to observe that in each task just a few annotators are assigned significantly larger weights than the majority, but each task had a few effectively adversarial annotators.

By comparing the mean inter-annotator agreements, it was found that the annotators assigned greater positive weights also tended to be more strongly correlated than those with non-positive weights. This was true for all three of the tasks, with the difference in mean weight between annotators assigned positive and negative correlation being 0.09, 0.2 and 0.32 for the classification, scoring and ranking tasks, respectively. These results suggest that optimising the accuracy on the gold standard also indirectly optimised the inter-annotator agreement.

7.3 Optimising Inter-Annotator Agreement

The results obtained by optimising the accuracy on the gold standard show that the consensus estimates can be improved by assigning unequal weights to the annotators. As noted, this optimisation also improved the inter-annotator agreement. To avoid using the gold standard instances, which might not always be available or, since they depend on only one or two "expert" annotators, may themselves be unreliable, we investigated weighting annotators to maximise the inter-annotator agreement. We adapted the optimisation procedure described above to maximise the inter-annotator agreement by assigning weights which increased the average Spearman rank correlation between annotators. Table 4 summarises the results from the optimisation, show-

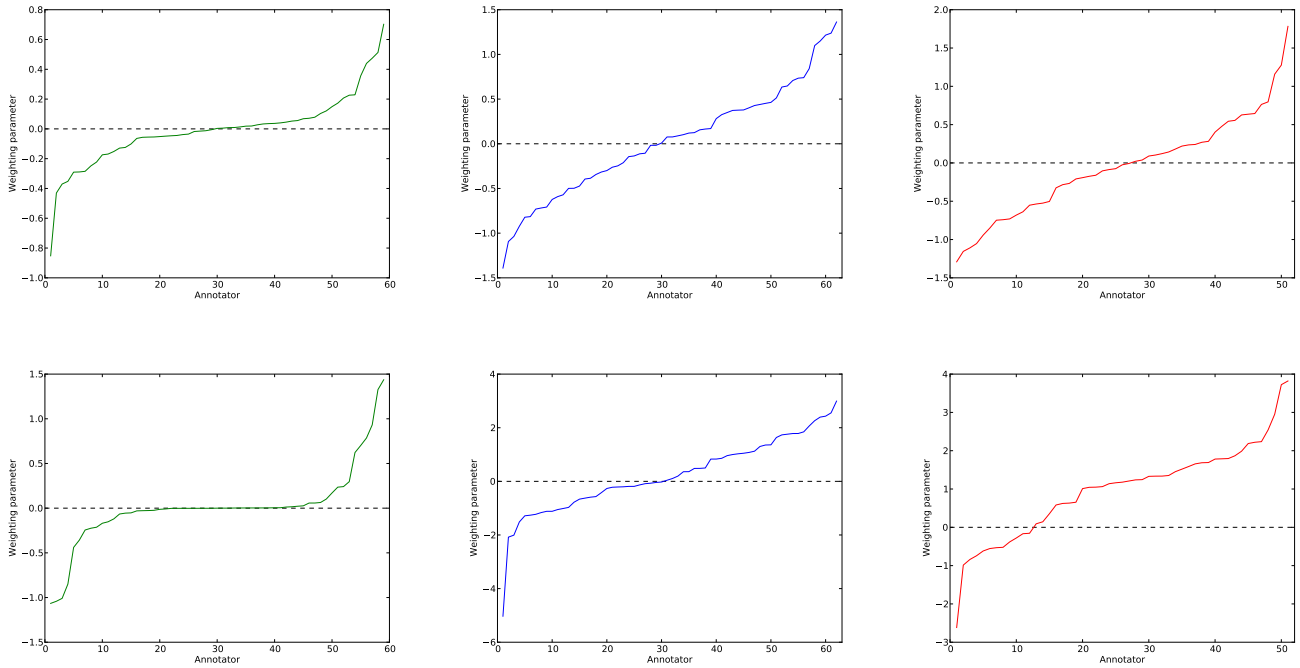


Fig. 9 The top row shows the optimised weighting parameters averaged over all LOO training runs for the classification, scoring and ranking tasks, respectively. The bottom row shows the weights obtained from optimising the inter-annotator agreement. The dashed horizontal line indicates zero weighting.

Table 4 Results from optimisation of the inter-annotator agreement. The table shows the average Spearman rank correlation obtained before and after optimisation.

Task	Orig ITA	Optim ITA
Classification	0.124	0.148
Scoring	0.166	0.211
Ranking	0.302	0.307

ing the inter-annotator agreement obtained both before and after optimisation. The value of λ was set to 0.075, 0.015 and 0.01 for the classification, scoring and ranking tasks. We also evaluated the accuracy of the annotators on the gold standard using the weights obtained from the optimisation. While the classification and ranking tasks maintained the same level of accuracy after optimisation, the scoring task increased from 0.57 to 0.71. Note that the inter-annotator agreement of the scoring task also increased the most during optimisation.

The second row of Figure 9 shows plots of the final optimised weighting parameters for each task. The results show that the classification task required a sparser set of weights in order to maximise the correlation. This corresponds to the weights obtained from the LOO optimisation, which also show more annotators assigned weights close to 0 for the classification task. We can interpret this as indicating the presence of more spammers in the population, which tend to decrease the accuracy and inter-annotator agreement due to their low correlation on average with the rest of the population. Assigning them low weight reduces their influence and enables a few higher quality annotators to dominate.

8 Conclusion

The use of citizen science for scoring annotations in microscopy images was shown to be viable. The annotations can be used to reliably characterise the degree of clumpiness within the images. They could also provide additional insights by indicating images with highly variable annotations, which may be due to anomalies within the cell. The results also demonstrate that it is possible for a relatively large number of image annotations to be obtained from a comparatively small number of non-expert annotators.

Although the tasks required significant effort from the participants due to the variation and complexity of the images, the accuracy of the annotators was still generally high. The annotator estimates on the gold standard compared favourably with the expert annotations overall. Annotators from each task were also shown to be consistent and reliable in their estimates, with those from the ranking task in particular showing a strong

similarity between the original and rotated image annotations. This is easily understood on recognising that annotators find it easier to order images rather than assigning images to an arbitrary scale or class, even when exemplars of the scale are available.

A significant improvement in accuracy can be obtained by optimising annotator weighting parameters. This was demonstrated using an evolutionary algorithm to improve the accuracy of the consensus estimates on the gold standard. Using the optimised parameters, the consensus for each task obtained a high level of accuracy on the gold standard. It was also found that annotators with higher mean inter-annotator agreement tended to be assigned greater weight, suggesting a correlation between the accuracy on the gold standard and the overall agreement among the annotators. By adapting the optimisation procedure to select weights which increase the correlation between annotators, it was shown that the inter-annotator agreement can also be optimised directly. For the scoring task, this also led to a significant improvement in accuracy on the gold standard.

The annotations obtained from the experiment demonstrate that the type of task presented to annotators had a significant impact on the quality of the resulting data. All three of the tasks showed clear differences in accuracy and inter-annotator agreement, with the ranking task obtaining the best overall performance.

References

R. A. Bailey. *Design of Comparative Experiments*. Cambridge University Press, Cambridge, UK, 1st edition, 2008.

A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM*, 54(4):86–96, 2011.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 0167-8655.

L. Fortson, K. Masters, R. Nichol, K. Borne, E. Edmondson, C. Lintott, J. Raddick, K. Schawinski, and J. Wallin. Galaxy Zoo: Morphological Classification and Citizen Science. In M. J. Way a d J. D. Scargle, K. M. Ali, and A. N. Srivastava, editors, *Advances in Machine Learning and Data Mining for Astronomy*, Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC, 2012.

A. Gelman, J. B. Carlin, H. S. Stren, D. B. Dunson, Vehtari A., and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2013.

J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212, New York, NY, 2010. ACM.

J. Knowles and D. Corne. The Pareto Archived Evolution Strategy: A New Baseline Algorithm for Pareto Multiobjective Optimisation. In *Proceedings of the Congress on Evolutionary Computation*, volume 1, pages 98–105, 1999.

G. Lebanon and J. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning*, pages 363–370, 2002.

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York, NY, 1st (revised) edition, 2006.

G.R. Littlejohn, J.D. Gouveia, C. Edner, N. Smirnoff, and J. Love. Perfluorodecalin substantially improves confocal depth resolution in air-filled tissues. *New Phytologist*, 186(4):1018–1025, 2010.

G. Parent and M. Eskenazi. Clustering dictionary definitions using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 21–29, Los Angeles, CA, June 2010. Association for Computational Linguistics.

V. C. Raykar and S. Yu. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research*, 13: 491–518, March 2012. ISSN 1532-4435.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11: 1297–1322, 2010.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the EMNLP Conference on Empirical Methods in Natural Language Processing*, pages 254–263, New York, NY, 2008. ACM.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267288, 1996.

W. Truman, M. de Torres Zabala, and M. Grant. A complex interplay of transcriptional regulation acts to modify basal defense responses during pathogenesis. *The Plant Journal*, 46:14–33, 2006.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043, 2009.